This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

PHA: Patch-wise High-frequency Augmentation for Transformer-based Person Re-identification

Guiwei Zhang[1], Yongfei Zhang[1,2,3]*, Tianyu Zhang[1], Bo Li[1,2], Shiliang Pu[4]
[1]Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University.
[2]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University.
[3] Pengcheng Laboratory. [4] Hikvision Research Institute.
{zhangguiwei,yfzhang,zhangtianyu,boli}@buaa.edu.cn, pushiliang.hri@hikvision.com

Abstract

*Although recent studies empirically show that injecting Convolutional Neural Networks (CNNs) into Vision Transformers (ViTs) can improve the performance of person re-identification, the rationale behind it remains elusive. From a frequency perspective, we reveal that ViTs perform worse than CNNs in preserving key high-frequency components (e.g, clothes texture details) since high-frequency components are inevitably diluted by low-frequency ones due to the intrinsic Self-Attention within ViTs. To remedy such inadequacy of the ViT, we propose a **P**atch-wise **H**igh-frequency **A**ugmentation (PHA) method with two core designs. **First**, to enhance the feature representation ability of high-frequency components, we split patches with high-frequency components by the Discrete Haar Wavelet Transform, then empower the ViT to take the split patches as auxiliary input. **Second**, to prevent high-frequency components from being diluted by low-frequency ones when taking the entire sequence as input during network optimization, we propose a novel patch-wise contrastive loss. From the view of gradient optimization, it acts as an implicit augmentation to improve the representation ability of key high-frequency components. This benefits the ViT to capture key high-frequency components to extract discriminative person representations. PHA is necessary during training and can be removed during inference, without bringing extra complexity. Extensive experiments on widely-used ReID datasets validate the effectiveness of our method.*

1. Introduction

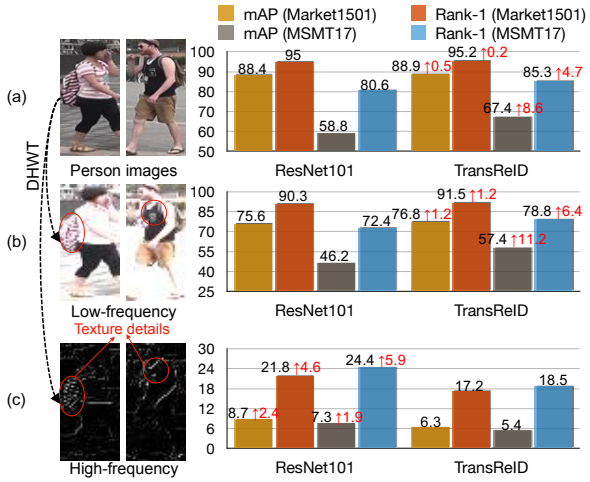

Figure 1. Performance comparisons between ResNet101 (**44.7M #Param**) and TransReID (**100M #Param**) on (a) original person images, (b) low-frequency components, and (c) high-frequency components of Market1501 and MSMT17 datasets, respectively. Note that "#Param" refers to the number of parameters.

Person re-identification (ReID) aims to retrieve a specific person, given in a query image, from the search over a large set of images captured by various cameras [33, 36, 37, 45]. Most research to date has focused on extracting discriminative person representations from single images, either by Convolutional Neural Networks (CNNs) [24, 25, 39, 42], Vision Transformers (ViTs) [2, 8, 23, 27, 41, 43] or Hybrid-based approaches [10, 11, 14, 16, 35]. These studies empirically show that injecting CNNs into ViTs can improve the discriminative of person representations [8, 35].

Despite a couple of empirical solutions, the rationale for why ViTs can be improved by CNNs remains elusive. To this end, we explore possible reasons from a frequency perspective, which is of great significance in digital image processing [4, 12, 32]. As shown in Fig. 1, we first employ Discrete Haar Wavelet Transform (DHWT) [19] to transform original person images into low-frequency compo-

*Corresponding author
https://github.com/zhangguiwei610/PHA

This work was partially supported by the National Natural Science Foundation of China (No. 62072022) and the Fundamental Research Funds for the Central Universities.

nents and high-frequency components, then conduct performance comparisons between the ResNet101 [6] and TransReID [8] on original images, the low-frequency components and high-frequency components of Market1501 and MSMT17 datasets, respectively. Our comparisons reveal:

(1) Certain texture details of person images, which are more related to the high-frequency components, are crucial for ReID tasks. Specifically, from Fig. 1 (a) and (b), with the same model, the performance on the original images is consistently better than that on low-frequency components. Taking the TransReID as an example, the Rank-1/mAP on original images of MSMT17 dataset is 6.5%/10.0% higher than that on the low-frequency components. The root reason might be that low-frequency components only reflect coarse-grained visual patterns of images, and lose texture details (e.g., bags and edges). In contrast, the lost details are more related to the high-frequency components, as shown in Fig. 1 (c). The degradation from Fig. 1 (a) to (b) indicates that certain details are key components to improve the performance of ReID.

(2) The ViT performs worse than CNNs in preserving key high-frequency components (e.g., texture details of clothes and bags) of person images. As shown in Fig. 1 (c), although the TransReID outperforms the ResNet101 consistently on the original images and low-frequency components, the ResNet101 exceeds the TransReID by 4.6%/2.4% and 5.9%/1.9% Rank-1/mAP on high-frequency components of Market1501 and MSMT17 datasets. The poor performance of the TransReID on high-frequency components shows its inadequacy in capturing key high-frequency details of person images.

In view of the above, we analyze the possible reason for such inadequacy of the ViT by revisiting Self-Attention from a frequency perspective (Sec. 3.1). We reveal that high-frequency components of person images are inevitably diluted by low-frequency ones due to the Self-Attention mechanism within ViTs. To remedy such inadequacy of the ViT without modifying its architecture, we propose a **P**atch-wise **H**igh-frequency **A**ugmentation (PHA) method with two core designs (Sec. 3.2). **First**, unlike previous works that directly take frequency subbands as network input [1, 4, 5, 17], we split patches with high-frequency components by the Discrete Haar Wavelet Transform (DHWT) and drop certain low-frequency components correspondingly, then empower the ViT to take the split patches as auxiliary input. This benefits the ViT to enhance the feature representation ability of high-frequency components. Note that the dropped components are imperceptible to human eyes but essential for the model, thereby preventing the model from overfitting to low-frequency components. **Second**, to prevent high-frequency components from being diluted by low-frequency ones when taking the entire sequence as input during network optimization, we propose a novel patch-wise contrastive loss. From the view of gradient optimization, it acts as an implicit augmentation to enhance the feature representation ability of key high-frequency components to extract discriminative person representations (Sec. 3.3). With it, our PHA is necessary during training and can be discarded during inference, without bringing extra complexity. Our contributions include:

- We reveal that due to the intrinsic Self-Attention mechanism, the ViT performs worse than CNNs in capturing high-frequency components of person images, which are key ingredients for ReID. Hence, we develop a **P**atch-wise **H**igh-frequency **A**ugmentation (PHA) to extract discriminative person representations by enhancing high-frequency components.

- We propose a novel patch-wise contrastive loss, enabling the ViT to preserve key high-frequency components of person images. From the view of gradient optimization, it acts as an implicit augmentation to enhance the feature representation ability of key high-frequency components. By virtue of it, our PHA is necessary during training and can be removed during inference, without bringing extra complexity.

- Extensive experimental results perform favorably against the mainstream methods on CUHK03-NP, Market-1501, and MSMT17 datasets.

2. Related Work

2.1. Transformer-based Person Re-identification

Recently, many works introduce the Vision Transformer (ViT) [2] into ReID tasks with great success, which can be roughly summarized into the following two aspects. (1) Pure ViT-based methods. He et al. [8] propose a pure ViT-based ReID framework, which outperforms the state-of-the-art CNN-based approaches. PASS [43] extracts fine-grained discriminative information with the ViT for person ReID. PFD [27] proposes a pose-guided Transformer encoder-decoder architecture for occluded person ReID tasks. DCAL [41] removes misleading attentions to extract complementary and discriminative part features. (2) Introducing CNNs into ViTs. PAT [16], DRL-Net [9] and APD [11] combine the CNN backbone and Transformer layers to extract discriminative parts representations. Wang et al [26] propose the Neighbor Transformer followed by a convolutional network to enhance interactions across all person images during training. Zhang et al propose the HAT [35], in which multi-scale features from CNNs are aggregated by Transformer blocks. Rather than modifying the model architecture, our PHA benefits the vanilla ViT to capture pivotal high-frequency components of person images, without bringing extra complexity during inference.

2.2. Application of Frequency Information in Vision

Frequency information is of great significance in digital image processing [4, 12]. Some works leverage the frequency information to improve the performance in vision tasks [1,5,29,31,32], while others accelerate the network in the frequency domain [20, 21]. Yang et al [31] transfer frequency components of images to improve domain adaptive semantic segmentation. Guo et al [5] takes low-frequency components of images as inputs to recover the missing details. Yao [32] et al leverage wavelet transform to downsample key/values within Transformer blocks, without information dropping. On the other hand, Rao et al [21] propose the GFNet to establish long-term spatial dependencies from a frequency perspective with only log-linear complexity. Oyallon [20] et al propose a wavelet scattering network, achieving comparable performance on image recognition with fewer parameters. Differently, our PHA benefits the ViT to extract discriminative person representations by enhancing key high-frequency components.

3. Method

We first analyze the possible reason for the above inadequacy of the ViT by revisiting Self-Attention from a frequency perspective (Sec. 3.1). Then, we propose the **P**atch-wise **H**igh-frequency **A**ugmentation (Sec. 3.2), including ❶ high-frequency enhancement and low-frequency drop, and ❷ the patch-wise contrastive loss. Finally, we analyze the effectiveness of the patch-wise contrastive loss from the view of gradient optimization (Sec. 3.3).

3.1. Analysis

Revisit Self-Attention. Given a person image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$, where *H, W and C* represent its height, width and the number of channels respectively, we first split it into N fixed-sized patches $\{\boldsymbol{x}_i | i = 1, 2, \cdots, N\}$. A learnable [cls] token denoted as $\boldsymbol{x}_{cls}$ is prepended to the input sequence. Learnable position embeddings $\mathcal{P} \in \mathbb{R}^{(1+N)\times D}$ are also applied to introduce spatial information. Hence, the input sequence fed into the ViT is formulated as below:

$$\boldsymbol{y} = [\boldsymbol{x}_{cls}; \mathcal{F}(\boldsymbol{x}_1); \mathcal{F}(\boldsymbol{x}_2); ..., \mathcal{F}(\boldsymbol{x}_N)] + \mathcal{P} \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^{(1+N)\times D}$ denotes the input sequence. $\mathcal{F}$ is a learnable projection that maps patches to D dimensions. Let $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{D\times D}$ denote the query, key and value projection matrices, respectively, the Self-Attention encodes each embedding in $\boldsymbol{y}$, formulated as below:

$$\boldsymbol{A}_{ij} = \frac{\boldsymbol{y}_i \boldsymbol{W}^Q (\boldsymbol{y}_j \boldsymbol{W}^K)^\top}{\sqrt{D}} \tag{2}$$

$$\boldsymbol{z}_i = \sum_j \sigma(\boldsymbol{A}_{ij}) \boldsymbol{y}_j \boldsymbol{W}^V \tag{3}$$

where $\sigma(\cdot)$ denotes the softmax function, $\boldsymbol{y}_i$ represents the i-th embedding in $\boldsymbol{y}$ and $\boldsymbol{z}_i$ is the corresponding output.

Analysis from a frequency perspective. One reason for such inadequacy may be that as the Transformer layer deepens, the high-frequency components are inevitably diluted by low-frequency ones due to Self-Attention. To verify it, we use the Discrete Haar Wavelet Transform (DHWT) to decompose a person image $\boldsymbol{x}$ into four wavelet subbands: $\boldsymbol{x}_{LL}$, $\boldsymbol{x}_{LH}$, $\boldsymbol{x}_{HL}$, and $\boldsymbol{x}_{HH} \in \mathbb{R}^{H/2\times W/2\times C}$. Note that $\boldsymbol{x}_{LL}$ denotes the low-frequency components that reflect the person structure at a coarse-grained level. We concatenate the later three subbands along the channel dimension:

$$\mathcal{M}_h(\boldsymbol{x}) = \textbf{concat}(\boldsymbol{x}_{LH}, \boldsymbol{x}_{HL}, \boldsymbol{x}_{HH}) \tag{4}$$

where $\mathcal{M}_h(\boldsymbol{x}) \in \mathbb{R}^{H/2\times W/2\times 3C}$ denotes the high-frequency components of images, which reflect fine-grained texture details. Then, we define set Ω, which contains the indices of patches with top-K high-frequency responses:

$$\Omega := \{j | j \in \text{top-}K(||\mathcal{G}(\mathcal{M}_h(\boldsymbol{x}))||_2)\} \tag{5}$$

where the function $\mathcal{G} : \mathbb{R}^{H/2\times W/2\times 3C} \to \mathbb{R}^{N\times 3C}$ includes the downsampling and flattening operations and $||\cdot||_2$ is the ℓ_2-norm. Hence, we rewrite Eq. (3) as a sum of two parts:

$$\boldsymbol{z}_i = \sum_{j\in\Omega} \sigma(\boldsymbol{A}_{ij}) \boldsymbol{y}_j \boldsymbol{W}^V + \sum_{j\notin\Omega} \sigma(\boldsymbol{A}_{ij}) \boldsymbol{y}_j \boldsymbol{W}^V \tag{6}$$

Assuming that the subscript i in Eq. (6) denotes an arbitrary index belonging to set Ω, the embedding $\boldsymbol{z}_i$ with high-frequency components is arguably a convex combination of embeddings with top-K high-frequency responses (**first part**) and the rest reflecting low-frequency components (**second part**). As the Transformer layer deepens, embeddings with high-frequency components are continuously affected by the significant presence of embeddings with low-frequency components in the current and previous layers. Although low-frequency components are critical, certain high-frequency ones, which are key ingredients for ReID, are inevitably diluted. To validate it, we define s as the indicator to evaluate the similarity between patches with top-K high-frequency responses and the rest:

$$\boldsymbol{p} = [p_{class}, p_1, p_2, ..., p_N], \boldsymbol{p} \in \mathbb{R}^{(1+N)\times D} \tag{7}$$

$$s = \frac{1}{K} \sum_{k\in\Omega} \frac{1}{N-K} \sum_{l\notin\Omega} \frac{|p_k^\top p_l|}{||p_k||_2 ||p_l||_2} \tag{8}$$

where $\boldsymbol{p}$ denotes the sequence of embeddings encoded when the entire sequence $\boldsymbol{y}$ is taken as input. Note that all embeddings except p_{class} in Eq. (7) participate in Eq. (8). Let $\boldsymbol{sim} = \frac{1}{T}\sum_{t=1} s_t$, where T is the number of all images

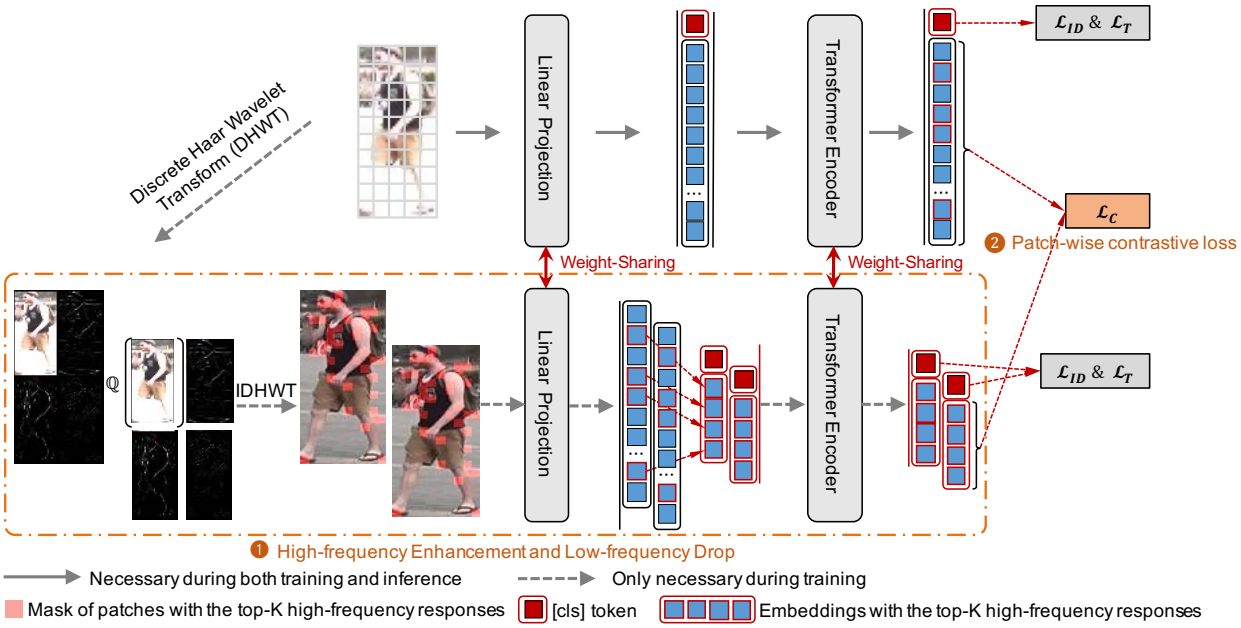

Figure 2. The overall of our proposed **P**atch-wise **H**igh-frequency **A**ugmentation (PHA), which consists of ❶ High-frequency enhancement and low-frequency drop, and ❷ a novel patch-wise contrastive loss. Note that "IDHWT" refers to Inverse Discrete Haar Wavelet Transform, and $\mathbb{Q}[\cdot]$ denotes the quantization operation. With PHA, the vanilla ViT could enhance the feature representation ability of pivotal high-frequency components. PHA is only necessary during training and can be removed during inference, without bringing extra complexity.

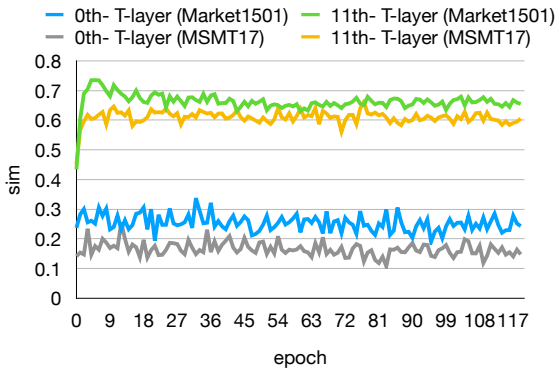

Figure 3. Comparisons in the indicator $\boldsymbol{sim}$ with "after the 0th- and 11th- Transformer layer" on Market1501 and MSMT17 datasets, respectively. Note that the "T-layer" refers to the Transformer layer for notation simplicity.

in the training set, Fig. 3 shows the comparisons in $\boldsymbol{sim}$ with "after the 0th- and 11th- Transformer layer" on two ReID datasets during training. Intuitively, the larger $\boldsymbol{sim}$ is, the more severe high-frequency components are smoothed. This is not beneficial to preserve key high-frequency components, which validates our analysis.

3.2. Patch-wise High-frequency Augmentation

To remedy such inadequacy of the ViT, we propose the **P**atch wise **H**igh-frequency **A**ugmentation (PHA) method. Fig. 2 shows the overall framework during training, which consists of ❶ High-frequency enhancement and low-frequency drop, and ❷ a novel patch-wise contrastive loss.

❶ High-frequency Enhancement and Low-frequency Drop. The purpose is to prevent key high-frequency components from being diluted by low-frequency components. Given the input sequence $\boldsymbol{y}$ depicted in Eq. (1), we first sample a subset of patches. The sampling strategy is straightforward: only the patches belonging to set Ω, which is depicted in Eq. (5), are sampled. Subsequently, we combine the [cls] token $\boldsymbol{x}_{cls}$ in Eq. (1) with the sampled subset to form the high-frequency sub-sequence $\boldsymbol{y}^h$. Then we empower the vanilla ViT to take $\boldsymbol{y}^h$ as auxiliary input, and the encoded output is formulated as below:

$$\boldsymbol{p}^h = \left[p^h_{class}, p^h_1, p^h_2, ..., p^h_K\right], \boldsymbol{p}^h \in \mathbb{R}^{(1+K)\times D} \quad (9)$$

To further prevent the ViT from overfitting to low-frequency components of person images, we drop certain low-frequency components by quantization:

$$\mathbb{Q}(\boldsymbol{x}_{LL}, q) = \left\lfloor \frac{\boldsymbol{x}_{LL} + 0.5}{q} \right\rfloor \cdot q \quad (10)$$

where q denotes the interval length, which determines the nearest quantization point for values of $\boldsymbol{x}_{LL}$. Intuitively, the larger of q, the more low-frequency components are dropped. Note that the dropped low-frequency components are imperceptible to human eyes but essential for the ViT

model. This is beneficial to prevent the model from overfitting to low-frequency components. Afterward, we apply inverse DHWT to reconstruct person images. Note that both original images and reconstructed images contribute to the high-frequency enhancement scheme.

Subsequently, the encoded representations p_{class} and p^h_{class} in Eq. (7) and Eq. (9) serve as the global person representation $\boldsymbol{f}_g$ and high-frequency enhanced representation $\boldsymbol{f}_h$. By convention, we optimize the representations $\boldsymbol{f}_g$ and $\boldsymbol{f}_h$ with the ID loss $\mathcal{L}_{ID}$ and the triplet loss $\mathcal{L}_T$:

$$\mathcal{L} = \mathcal{L}_{ID}(\boldsymbol{f}_g) + \mathcal{L}_T(\boldsymbol{f}_g) + \mathcal{L}_{ID}(\boldsymbol{f}_h) + \mathcal{L}_T(\boldsymbol{f}_h) \quad (11)$$

This benefits the ViT to enhance the representation ability of pivotal high-frequency components, e.g., clothes texture, to extract discriminative person representations.

❷Patch-wise Contrastive Loss. To prevent key high-frequency components from being over-smoothed by low-frequency ones when the ViT takes the entire sequence as input during network optimization, we propose a novel patch-wise contrastive loss. Let $\boldsymbol{P} \in \mathbb{R}^{B\times(1+N)\times D}$ and $\boldsymbol{P}^h \in \mathbb{R}^{B\times(1+K)\times D}$ correspond to the encoded entire sequence (Eq. (7)) and the high-frequency enhanced subsequence (Eq. (9)) in each mini-batch. We aim to pull the embeddings in $\boldsymbol{P}$, which belong to set Ω, closer together with high-frequency enhanced embeddings from the same identity in $\boldsymbol{P}^h$, while pushing the embeddings from different identities apart. Specifically, we have:

$$S(i,j,k) = \exp(\frac{\boldsymbol{P}_{i,k}}{||\boldsymbol{P}_{i,k}||_2}^{\top} \cdot \frac{\boldsymbol{P}^h_{j,k}}{||\boldsymbol{P}^h_{j,k}||_2}) \quad (12)$$

$$\mathcal{L}_C = -\frac{1}{K}\sum_{\underline{k\in\Omega}}\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{1}{M}\sum_{j:l_j=l_i}S(i,j,k) - \log\left(\frac{1}{M}\sum_{j:l_j=l_i}S(i,j,k) + \frac{\tau}{B-M}\sum_{j:l_j\neq l_i}S(i,j,k)\right)\right] \quad (13)$$

where M is the number of images per identity in each mini-batch, l_i is the label of the i-th image, and τ denotes the temperature factor. Please note that we only regularize the patches in $\boldsymbol{P}$ that belong to set Ω. (underlined in Eq. (13) for emphasis).

By virtue of it, the ViT could enhance the feature representation ability of embeddings with top-K high-frequency responses even when taking the entire sequence $\boldsymbol{y}$ as input. This benefits the ViT to preserve key high-frequency components (e.g., texture details of clothes) of person images, to extract discriminative person representations.

3.3. Why the Contrastive Loss Works?

We analyze the effectiveness of **P**atch-wise **C**ontrastive **L**oss (PCL) $\mathcal{L}_C$ from the perspective of gradient optimization. Specifically, the PCL propagates new gradients on high-frequency enhanced embeddings in $\boldsymbol{P}^h$. As illustrated in Fig. 4, given an arbitrary spatial position k belonging to set Ω, let $\widetilde{\boldsymbol{P}}_k = \{\boldsymbol{P}_{i,k}|i = 1,2,...,B\}$ and $\widetilde{\boldsymbol{P}}^h_k = \{\boldsymbol{P}^h_{i,k}|i = 1,2,...,B\}$ represent the corresponding patch embeddings in $\boldsymbol{P}$ and $\boldsymbol{P}^h$, and $\widetilde{\boldsymbol{\mathcal{L}}}_k = \{\mathcal{L}_{k,i}|i = 1,2,...,B\}$ denotes the corresponding contrastive loss. Then, we have:

$$\frac{\partial\widetilde{\boldsymbol{\mathcal{L}}}_k}{\partial\widetilde{\boldsymbol{P}}^h_k} = \sum_i^B\left(\frac{\partial\mathcal{L}_{k,i}}{\partial\boldsymbol{P}^h_{i,k}} + \sum_{j\neq i}^B\frac{\partial\mathcal{L}_{k,i}}{\partial\boldsymbol{P}^h_{j,k}}\right) \quad (14)$$

Eq. (14) shows that the contrastive loss propagates B kinds of different gradient items on each high-frequency enhanced embedding in $\widetilde{\boldsymbol{P}}^h_k$. In contrast, without the contrastive loss, all losses only propagate gradients on the [cls] token. We emphasize that the patch-wise contrastive loss can be viewed as an implicit augmentation that brings each embedding in $\widetilde{\boldsymbol{P}}^h_k$ close to embeddings from the same identity in $\widetilde{\boldsymbol{P}}_k$. This is beneficial to prevent the network from overfitting to identity-irrelevant high-frequency components, while enhancing the feature representation ability of key high-frequency components.

Finally, we optimize our PHA method by minimizing the overall objective with identity labels:

$$\mathcal{L}_{overall} = \mathcal{L} + \mathcal{L}_C \quad (15)$$

Note that our proposed PHA method is only necessary during training and can be discarded during inference, without bringing extra complexity.

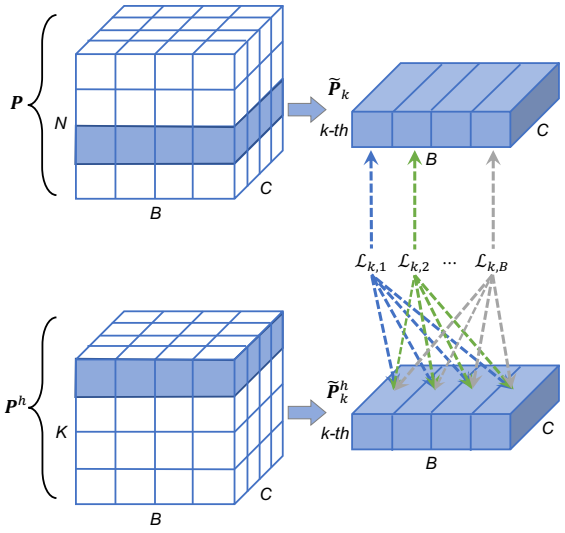

Figure 4. Illustration of the patch-wise contrastive loss. Dashed lines denote B kinds of different gradient propagations among each high-frequency enhanced embedding in $\widetilde{\boldsymbol{P}}^h_k$. The [cls] tokens in $\boldsymbol{P}$ and $\boldsymbol{P}^h$ are omitted from the figure for simplicity.

Method	Market1501		MSMT17		CUHK03-NP			
					Labeled		Detected	
	R1(%)	mAP(%)	R1(%)	mAP(%)	R1(%)	mAP(%)	R1(%)	mAP(%)
CNN-based methods								
STF (ICCV 19) [18]	93.4	82.7	73.6	47.6	68.2	62.4	-	-
BAT-net (ICCV 19) [3]	94.1	85.5	79.5	56.8	78.6	76.1	76.2	73.2
ISP (ECCV 20) [42]	95.3	88.6	-	-	76.5	74.1	75.2	71.4
RGA-SC (CVPR 20) [38]	96.1	88.4	80.3	57.5	81.1	77.4	79.6	74.5
CBN (ECCV 20) [45]	94.3	83.6	72.8	42.9	-	-	-	-
CBDB-Net(TCSVT 21) [22]	94.4	85.0	-	-	77.8	76.6	75.4	72.8
CDNet (CVPR 21) [13]	95.1	86.0	78.9	54.7	-	-	-	-
C2F (CVPR 21) [34]	94.8	87.7	-	-	80.6	79.3	<u>81.3</u>	**84.1**
ViT-based methods								
DRL-Net (TMM 21) [9]	94.7	86.9	78.4	55.3	-	-	-	-
AAformer (arXiv 21) [44]	95.4	87.7	63.2	83.6	79.9	77.8	77.6	74.8
HAT (ACM 21) [35]	95.6	89.5	82.3	61.2	<u>82.6</u>	<u>80.0</u>		
TransReID (ICCV 21) [8]	95.2	88.9	<u>85.3</u>	<u>67.4</u>	81.7	79.6	79.6	77.0
PFD (AAAI 22) [27]	95.5	89.7	83.8	64.4	-	-	-	-
ABDNet+NFormer (CVPR 22) [26]	<u>95.7</u>	**93.0**	80.8	62.2	80.6	79.1	79.0	76.4
DCAL (CVPR 22) [41]	94.7	87.5	83.1	64.0	-	-	-	-
TransReID + PHA	**96.1**	<u>90.2</u>	**86.1**	**68.9**	**84.5**	**83.0**	**83.2**	<u>80.3</u>

Table 1. Comparison with the state-of-the-art models on Market-1501, MSMT17, and CUHK03-NP datasets. R1 means Rank-1 accuracy. Optimal and suboptimal results are highlighted in bold and underlined, respectively.

Dataset	ID	image	cams
MSMT17	4101	126441	15
Market-1501	1501	32668	6
CUHK03-NP	1467	13164	2

Table 2. Statistics of datasets used in the paper.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct extensive experiments on three standard person ReID benchmarks: Market-1501 [40], CUHK03-NP [15] and MSMT17 [28]. Table 2 shows details of above datasets. Following conventions in the ReID community [7, 8, 30], we adopt Cumulative Matching Characteristic (CMC) curves and the mean Average Precision (mAP) to evaluate the quality of different methods.

4.2. Implementation Details

Following TransReID [8], all input images are resized to 256×128 and the training images are augmented with random horizontal flipping, padding, random cropping and random erasing. The batch size is set to 64 with 4 images per ID and SGD optimizer is employed with a momentum of 0.9 and the weight decay of 0.0001. The learning rate is initialized as 0.008 with cosine learning rate decay. The parameter K in Eq. (5) is set to 35% and the interval length in Eq. (10) is set to 5. All experiments are performed with one Nvidia V100 GPU with FP16 training.

4.3. Comparison with the State-of-the-Art

We compare our method with the state-of-the-art approaches on widely-used person ReID datasets in Table 1. Our method achieves competitive performance compared to the prior CNNs-based and ViTs-based methods. Particularly, with the TransReID baseline, our method achieves 96.1%/90.2%, 86.1%/68.9%, 84.5%/83.0%, 83.2%/80.3% Rank-1/mAP on Market1501, MSMT17, CUHK03-NP labeled and CUHK03-NP detected datasets, respectively.

Comparison to ViT-based Methods. Some typical works (e.g., PAT [16], PFD [27] and DCAL [41]), extract discriminative part features for accurate alignment. Rather than aligning fine-grained parts, our PHA method benefits the ViT to preserve pivotal high-frequency components of images, to extract discriminative person representations. Compared to HAT [35] which aggregates hierarchical features from CNN with Transformer blocks, our PHA method does not modify the model architecture. It is only necessary during training and can be discarded during inference, without bringing extra computation costs.

Comparison to CNN-based Methods. Compared with the competing method C2F [34], our PHA outperforms it by 1.3%/2.5% and 3.9%/3.7% Rank-1/mAP on Market1501 and CUHK03-NP labeled datasets when taking the TransReID as the baseline. By virtue of our PHA, the ViT could not only build long-distance dependencies of low-frequency components but also capture key high-frequency components of person images. This benefits the ViT to extract discriminative person representations.

Index	HE	LD	PCL	CUHK03-NP R1 (%)	CUHK03-NP mAP (%)
1				81.7	79.6
2	✓			82.9	80.7
3	✓	✓		83.3	81.7
4	✓		✓	83.9	82.5
5		✓	✓	83.0	81.1
6	✓	✓	✓	84.5	83.0

Table 3. Ablation study over CUHK03-NP labeled dataset.

Method	Market-1501 R1 (%)	Market-1501 mAP (%)	CUHK03-NP R1 (%)	CUHK03-NP mAP (%)
w/ stop gradient	95.6	89.5	83.6	82.1
w/o stop gradient	**96.1**	**90.2**	**84.5**	**83.0**

Table 4. Comparison with "with *vs* without stop gradient on $\mathcal{L}_C$" on Market-1501 and CUHK03-NP labeled datasets.

4.4. Ablation Study

We conduct ablation studies on CUHK03-NP labeled dataset to analyze each core design, including **H**igh-frequency **E**nhancement (HE), **L**ow-frequency **D**rop (LD), and the **P**atch-wise **C**ontrastive **L**oss (PCL). We consider TransReID as the Baseline, and results are shown in Table 3.

Effectiveness of High-frequency Enhancement. From index-1 and index-2, when High-frequency Enhancement is applied, the Rank-1/mAP is greatly improved by 1.2%/1.1%. From index-5 and index-6, the Rank-1/mAP is further improved by 1.5%/1.9%. The results indicate that High-frequency Enhancement plays a vital role in enhancing the feature representation ability and discrimination of key high-frequency components (e.g, clothes texture).

Effectiveness of Low-frequency Drop. From index-2 and index-3, the Low-frequency Drop can improve the Rank-1/mAP by 0.4%/1.0%. From index-4 and index-6, the Rank-1/mAP is further improved by 0.6%/0.5%. The results show that the dropped low-frequency components, although imperceptible to human eyes, are beneficial to prevent ViT from overfitting the low-frequency components. Moreover, Low-frequency Drop complements High-frequency Enhancement to enhance the representation ability of key high-frequency components.

Effectiveness of Patch-wise Contrastive Loss. From index-2 and index-4, the Rank-1/mAP is improved by 1.0%/1.8%. From index-3 and index-6, the Rank-1/mAP is further improved by 1.2%/1.3%. The improvements indicate that the PCL benefits the ViT to enhance the feature representation ability of key high-frequency components, to extract discriminative person representations.

To verify that PCL works as an implicit augmentation, we further conduct a comparison on "with *vs.* without stop-gradient (stopgrad)" on high-frequency enhanced embeddings in $\boldsymbol{P}_k^h$. We implement it by modifying Eq. (12) as:

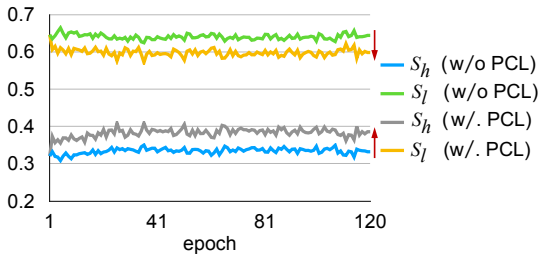

Figure 5. Comparisons in S_h and S_l between "with *vs.* without PCL" on CUHK03-NP labeled dataset.

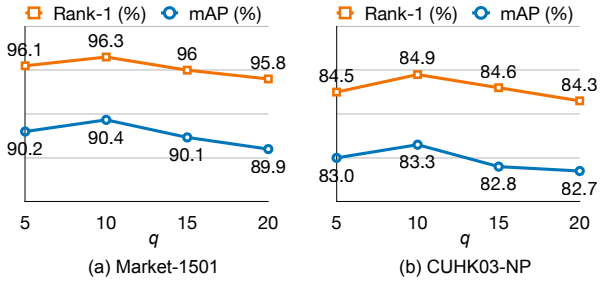

Figure 6. Comparison in Rank-1/mAP with different q on (a) Market-1501 and (b) CUHK03-NP labeled datasets.

$$S(i,j,k) = \exp\left(\frac{\boldsymbol{P}_{i,k}}{||\boldsymbol{P}_{i,k}||_2}^{\top} \cdot \text{stopgrad}\left(\frac{\boldsymbol{P}_{j,k}^h}{||\boldsymbol{P}_{j,k}^h||_2}\right)\right) \quad (16)$$

This means that the enhanced embedding in $\boldsymbol{P}_k^h$ is treated as a constant without gradient items propagating over it, and hence cannot be implicitly augmented by embeddings from $\boldsymbol{P}_k$. The results in Tab. 4 validate that propagating new gradient terms over $\boldsymbol{P}_k^h$ is of great significance in elevating the representation ability of key high-frequency components.

Another important contribution of PCL is to effectively balance the effects of high-frequency and low-frequency components. To explain it intuitively, we compute the row average of the attention matrix $\frac{1}{L}\sum_{i=1}^{L}\boldsymbol{A}_i$, where L is the number of ViT layers, then generate an attention vector $\boldsymbol{a} \in \mathbb{R}^{1+N}$. Since each row vector of $\boldsymbol{A}_i$ reveals how much each input embedding influences the resulting embeddings, we consider each value in $\boldsymbol{a}$ as the contribution of the embedding. We design two scores $S_h = \sum_{i\in\Omega}\boldsymbol{a}_i$ and $S_l = \sum_{j\notin\Omega}\boldsymbol{a}_j$ to evaluate the contribution of embeddings with high-frequency responses and the rest reflecting low-frequency ones, respectively. In Fig. 5, PCL amplifies high-frequency components and effectively adjusts the contribution of low-frequency ones, validating our analysis.

The Impact of interval length q in Eq. (10). We further conduct experiments to explore the most suitable q. As shown in Fig. 6, when q is set to 10, the Rank-1/mAP accuracy achieves the best. When q exceeds 10, the performance continuously decays. One possible reason might be that a too-high value of q falsely dropped key low-frequency components, which are also crucial for ReID tasks.

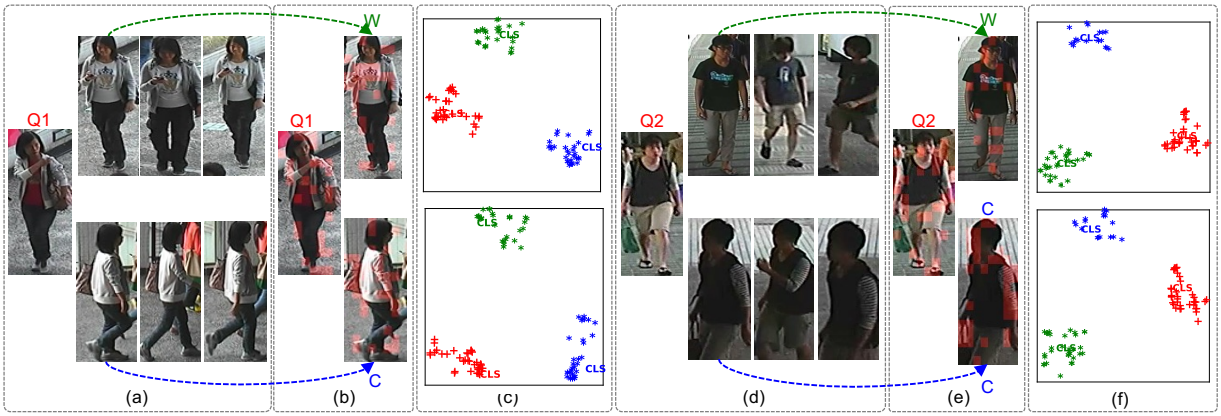

Figure 7. Comparison in top-3 ranking results and t-SNE visualization between the Baseline and our PHA method. In (a) and (d), the 1st- and 2nd- rows show the top-3 ranking results of the Baseline and our PHA method, respectively. The top-3 ranking results of the Baseline are wrong, while the results of our PHA are correct. In (b) and (e), we apply DHWT to find high-frequency patches of the query image "Q1"/"Q2", the wrong top-1 result "W", and the correct result "C". In (c) and (f), we conduct a comparison in t-SNE visualization of embeddings from "Q1"/"Q2", "W" and "C" between the Baseline (**top plot**) and our PHA method (**bottom plot**). To intuitively visualize the distribution of patch embeddings with high-frequency components, the low-frequency patch embeddings are omitted from the plot. Note that the symbol "CLS" in the plot denotes the constructed person representation.

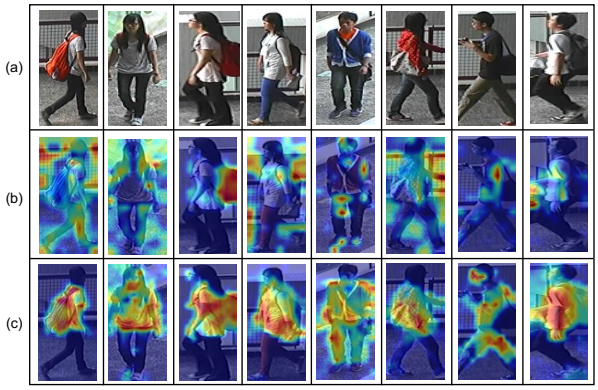

Figure 8. Grad-CAM visualization of attention maps **in inference**. (a) original person images. (b) attention maps of the Baseline. (c) attention maps of our PHA method, which could generate higher responses on key high-frequency ingredients.

4.5. Visualization

Top-3 ranking results and t-SNE visualization. Fig. 7 exhibits comparisons in top-3 ranking results and t-SNE visualization between the Baseline and our PHA method. Taking the query image "Q1" in Fig. 7 (a) as an example, the 1st- and 2nd- rows show the top-3 ranking results of the Baseline (wrong) and our method (correct), respectively. To explain the effectiveness of our PHA against the Baseline, we first apply DHWT to find high-frequency patches of the query image "Q1", the wrong top-1 result "W", and the correct result "C", respectively, as shown in Fig. 7 (b). In Fig. 7 (c), we conduct a comparison in t-SNE visualization of embeddings from "Q1", "W" and "C" between the Baseline (**top plot**) and our PHA (**bottom plot**). To intuitively visualize the distribution of patch embeddings with high-frequency components, the low-frequency patches are omitted from the plot. Note that the symbol "CLS" in the plot denotes the constructed person representations. Obviously, our PHA preserves a more diverse distribution of high-frequency patches, hence enabling the ViT to capture key high-frequency ingredients, e.g., the texture details of the red shirt and white shirt in "Q1" and "W". By virtue of it, the constructed person representations from different identities are pushed apart, while the representations from the same identity are pulled together.

Grad-CAM visualization of attention maps. Fig. 8 exhibits the attention maps of the Baseline and our PHA **during inference**. It can be seen that the Baseline performs poorly in capturing key high-frequency ingredients (e.g., bags, heads, and texture details of clothes). In contrast, our PHA could generate higher responses on the above details. This benefits the ViT to preserve key high-frequency components to extract discriminative person representations.

5. Conclusion

In this work, we reveal that the ViT perform worse than CNNs in preserving high-frequency components of person images. To remedy such inadequacy, we developed a **P**atch-wise **H**igh-frequency **A**ugmentation (PHA) method. To elevate the feature representation ability of high-frequency components when taking the entire sequence as input, a novel patch-wise contrastive loss is proposed. Note that the PHA is only necessary during training, without bringing extra complexity during inference. Extensive experimental results outperform almost all kinds of methods.

References

[1] Woong Bae, Jaejun Yoo, and Jong Chul Ye. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 145–153, 2017. 2, 3

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2

[3] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8030–8039, 2019. 6

[4] Shin Fujieda, Kohei Takayama, and Toshiya Hachisuka. Wavelet convolutional neural networks. *arXiv preprint arXiv:1805.08620*, 2018. 1, 2, 3

[5] Tiantong Guo, Hojjat Seyed Mousavi, Tiep Huu Vu, and Vishal Monga. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 104–113, 2017. 2, 3

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[7] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 6

[8] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15013–15022, 2021. 1, 2, 6

[9] Mengxi Jia, Xinhua Cheng, Shijian Lu, and Jian Zhang. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Transactions on Multimedia*, 2022. 2, 6

[10] Mengxi Jia, Xinhua Cheng, Yunpeng Zhai, Shijian Lu, Siwei Ma, Yonghong Tian, and Jian Zhang. Matching on sets: Conquer occluded person re-identification without alignment. In *Proc. AAAI Conf. Artif. Intell.*, pages 1673–1681, 2021. 1

[11] Shenqi Lai, Zhenhua Chai, and Xiaolin Wei. Transformer meets part model: Adaptive part division for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4150–4157, 2021. 1, 2

[12] Jae-Han Lee, Minhyeok Heo, Kyung-Rae Kim, and Chang-Su Kim. Single-image depth estimation based on fourier domain analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 330–339, 2018. 1, 3

[13] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6729–6738, 2021. 6

[14] He Li, Mang Ye, Cong Wang, and Bo Du. Pyramidal transformer with conv-patchify for person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7317–7326, 2022. 1

[15] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 6

[16] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021. 1, 2, 6

[17] Pengju Liu, Hongzhi Zhang, Kai Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 773–782, 2018. 2

[18] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4976–4985, 2019. 6

[19] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989. 1

[20] Edouard Oyallon, Eugene Belilovsky, and Sergey Zagoruyko. Scaling the scattering transform: Deep hybrid networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5618–5627, 2017. 3

[21] Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global filter networks for image classification. *Advances in Neural Information Processing Systems*, 34:980–993, 2021. 3

[22] Hongchen Tan, Xiuping Liu, Yuhao Bian, Huasheng Wang, and Baocai Yin. Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):160–171, 2021. 6

[23] Lei Tan, Pingyang Dai, Rongrong Ji, and Yongjian Wu. Dynamic prototype mask for occluded person re-identification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 531–540, 2022. 1

[24] Guangcong Wang, Jian-Huang Lai, Wenqi Liang, and Guangrun Wang. Smoothing adversarial domain attack and p-memory reconsolidation for cross-domain person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10568–10577, 2020. 1

[25] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the*

26th ACM international conference on Multimedia, pages 274–282, 2018. 1

[26] Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. Nformer: Robust person re-identification with neighbor transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7297–7307, 2022. 2, 6

[27] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. *arXiv preprint arXiv:2112.02466*, 2021. 1, 2, 6

[28] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 6

[29] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 3

[30] Cheng Yan, Guansong Pang, Xiao Bai, Changhong Liu, Ning Xin, Lin Gu, and Jun Zhou. Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss. *IEEE Transactions on Multimedia*, 2021. 6

[31] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 3

[32] Ting Yao, Yingwei Pan, Yehao Li, Chong-Wah Ngo, and Tao Mei. Wave-vit: Unifying wavelet and transformers for visual representation learning. *arXiv preprint arXiv:2207.04978*, 2022. 1, 3

[33] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[34] Anguo Zhang, Yueming Gao, Yuzhen Niu, Wenxi Liu, and Yongcheng Zhou. Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 598–607, 2021. 6

[35] Guowen Zhang, Pingping Zhang, Jinqing Qi, and Huchuan Lu. Hat: Hierarchical aggregation transformers for person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 516–525, 2021. 1, 2, 6

[36] Tianyu Zhang, Lingxi Xie, Longhui Wei, Yongfei Zhang, Bo Li, and Qi Tian. Single camera training for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12878–12885, 2020. 1

[37] Tianyu Zhang, Lingxi Xie, Longhui Wei, Zijie Zhuang, Yongfei Zhang, Bo Li, and Qi Tian. Unrealperson: An adaptive pipeline towards costless person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11506–11515, 2021. 1

[38] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020. 6

[39] Kecheng Zheng, Cuiling Lan, Wenjun Zeng, Jiawei Liu, Zhizheng Zhang, and Zheng-Jun Zha. Pose-guided feature learning with knowledge distillation for occluded person re-identification. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4537–4545, 2021. 1

[40] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 6

[41] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4692–4702, 2022. 1, 2, 6

[42] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer, 2020. 1, 6

[43] Kuan Zhu, Haiyun Guo, Tianyi Yan, Yousong Zhu, Jinqiao Wang, and Ming Tang. Part-aware self-supervised pre-training for person re-identification. *arXiv preprint arXiv:2203.03931*, 2022. 1, 2

[44] Kuan Zhu, Haiyun Guo, Shiliang Zhang, Yaowei Wang, Gaopan Huang, Honglin Qiao, Jing Liu, Jinqiao Wang, and Ming Tang. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*, 2021. 6

[45] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *European Conference on Computer Vision*, pages 140–157. Springer, 2020. 1, 6